

---

# Large Language Models, and LLM-Based Agents, Should Be Used to Enhance the Digital Public Sphere

---

Seth Lazar   Luke Thorburn   Tian Jin   Luca Belli  
ANU            KCL            MIT            Sator Labs

July 2, 2025

## Abstract

This paper argues that large language model-based recommenders can displace today’s attention-allocation machinery. LLM-based recommenders would ingest open-web content, infer a user’s natural-language goals, and present information that matches their reflective preferences. Properly designed, they could deliver personalization without industrial-scale data hoarding, return control to individuals, optimize for genuine ends rather than click-through proxies, and support autonomous attention management. Synthesizing evidence of current systems’ harms with recent work on LLM-driven pipelines, we identify four key research hurdles: generating candidates without centralized data, maintaining computational efficiency, modeling preferences robustly, and defending against prompt-injection. None looks prohibitive; surmounting them would steer the digital public sphere toward democratic, human-centered values.

## 1 Introduction

The shortcomings of contemporary online communication are well documented [15, 38]. As social media ecosystems splinter, indicators of abuse, epistemic pollution, manipulation and polarization continue to rise [91]. Consequently, the digital public sphere falls well below the optimistic projections of the early ‘wealth of networks’ thesis [3, 15, 55].

The causes of these problems are contested. Some scholars attribute them to human communicative behavior [90, 1]; others implicate the technical architectures that mediate online interaction. A decisive answer is improbable. Yet, while large-scale behavioral change is elusive, technical systems can be redesigned.

In this paper, we therefore concentrate on the technologies that allocate online attention. Regardless of whether they are the root cause of current pathologies, these systems are entangled with practices that are independently objectionable. This observation motivates a search for alternatives.

**We contend that prevailing recommender pipelines rely on mass surveillance, reinforce concentrated platform power, embody narrow behaviorist**

**assumptions, and erode user agency.** We advocate an alternative paradigm that leverages advances in large language models (LLMs) and LLM-based agents.<sup>1</sup> By representing both content and user values in natural language and invoking external tools when needed, such systems could avoid these four pathologies.

Avoiding these harms will not guarantee a healthier digital public sphere. If underlying social dynamics drive toxicity, improved recommendation may have limited effect, and commercial incentives could push practitioners to graft LLM recommenders onto existing surveillance-based models [78, 47, 48, 79]. Nevertheless, LLM-based recommenders can be deployed outside incumbent platforms, offering researchers and developers a plausible route around current gatekeepers. We therefore propose that technical researchers motivated to improve the digital public sphere should direct their attention to this approach.

Section 2 surveys the current state of online communication. Section 3 analyses how mainstream recommenders enable the four harms above. Section 4 reviews incremental responses and introduces LLM-based recommendation. Section 5 explains how LLM systems might mitigate the identified problems. Section 6 outlines outstanding (but surmountable) technical and ethical challenges; Section 7 concludes.

## 2 Diagnosing the Digital Public Sphere

The digital public sphere—the social, cultural, and political exchanges hosted online—can be evaluated along two axes. First, by the harms it enables; second, by the positive ideals it fails to achieve.

**Harms.** Following Lazar [55], these harms cluster into three categories

- *Abuse*: direct and indirect harassment, silencing, and other forms of targeted hostility [107, 86].
- *Epistemic pollution*: practices that obstruct accurate belief-formation—mis- and disinformation, coordinated inauthentic behavior, ‘flood-the-zone’ tactics, SEO-driven content farms, generative-AI ‘slop,’ and the ‘liar’s dividend’ [36, 106, 46, 42].
- *Manipulation*: communicative strategies that compromise agency, ranging from individual grooming and radicalization to population-level shifts such as affective polarization or body-image anxiety [93, 4, 84, 16].

**Positive ideals.** A healthy public sphere is not merely the absence of pathology; it realizes communicative goods—equitably satisfying people’s interests in sharing, receiving, and deliberating information. We label this objective *communicative justice* [55]. Designing for justice therefore requires technologies that advance positive aims as well as mitigate harm.

Attention in the digital public sphere is currently allocated by deep reinforcement learning-based recommender systems [67, 96, 68]. On large platforms (e.g., Facebook, Instagram, YouTube, X, TikTok), recommender pipelines typically:

- Retrieve a candidate set via lightweight heuristics.
- Rank candidates with a computationally heavier value model—usually a weighted sum of predicted engagement metrics, survey-based quality signals, and policy classifier scores.
- Re-rank to enforce secondary constraints (e.g., diversity across consecutive items, or fairness to producers, as in Spotify or YouTube—[65, 66, 88, 100, 9, 6]).

---

<sup>1</sup> Note that we are using ‘LLM’ here to mean any transformer-based token-sequence predictor, including multimodal foundation models, vision-language models, and others.

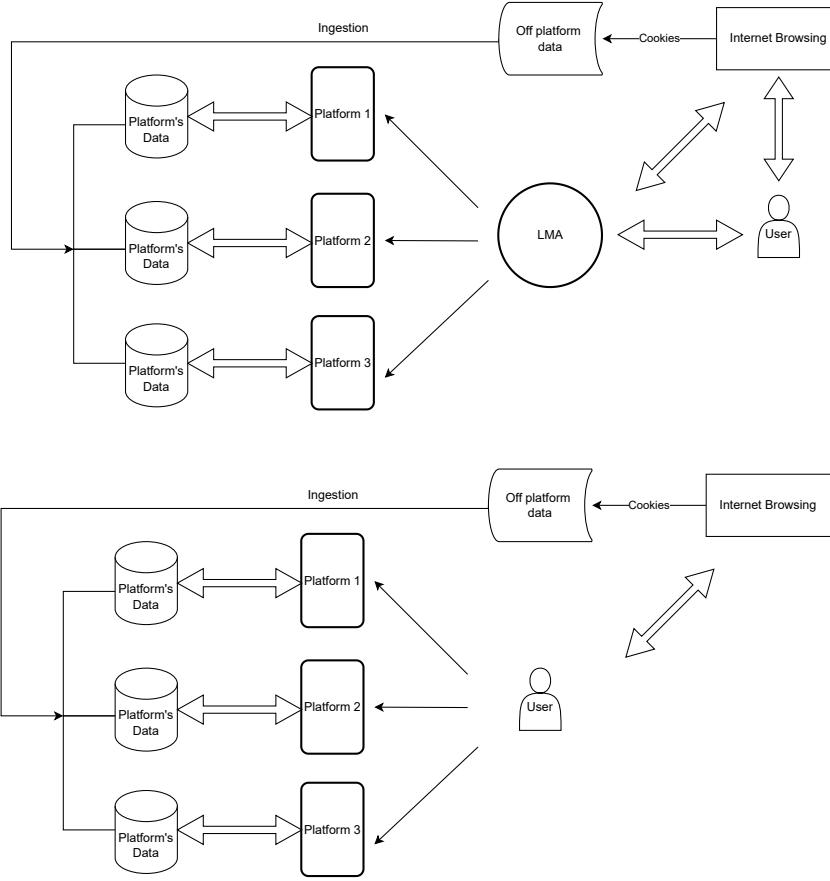


Figure 1: A schematic showing the core idea of using a Language Model Agent (LMA) as a recommender (top), compared to the current paradigm (below). In contrast to the current paradigm, an LMA could interact with web content (including but not only social media content) to source candidate items to recommend.

Because recommenders mediate much of the digital public sphere, they are attractive scapegoats for its shortcomings [25, 80]. Yet rigorous evidence for their causal role remains mixed [70, 33, 37, 72]. Decades of scholarship in science and technology studies (STS) also caution against technological determinism—the thesis that a technology alone can strictly determine significant societal outcomes. In general, societal impacts derive from complex interactions among technology, users, and social context [20].

## 2.1 A Narrow Target

Our aim is therefore limited. We identify four structural failings that today’s recommender paradigm tends to reinforce—**mass surveillance**, **power concentration**, **narrow behaviorism**, and **diminished user agency**—and argue that recommender systems built around LMs, especially agentic variants, could be designed to perform the same function while avoiding these specific pathologies. Crucially, such systems need not depend on the cooperation of incumbent gatekeepers. Even if their deployment leaves other problems of the digital public sphere untouched, reducing surveillance, countering concentration, broadening behavioral assumptions, and restoring user agency are worth

attempting for their own sake.

### 3 Four Problems with Recommender Systems

#### 3.1 Dependence on Mass Surveillance

At a functional level, any recommender must (1) represent the items competing for a user’s attention, (2) model the user’s preferences and values, and (3) select and order items so as to serve those preferences. The dominant industrial approach satisfies all three requirements by observing user behavior at scale:

- **Content understanding.** Recommenders infer an item’s salient properties from how large numbers of users interact with it.
- **User modeling.** They characterize a given user by comparing that user’s past engagement to aggregated behavioral profiles of others.
- **Inference and ranking.** They combine both behavioral streams to predict which items will maximize a proprietary ‘value function’—typically a weighted sum of engagement metrics, predicted surveys, and policy classifier scores.

Without continuous, fine-grained logging of impressions, clicks, likes, and dwell time, these systems degrade sharply. This was true for early click-through-rate pipelines that shaped Facebook’s ad placement and News Feed [34, 39]. It remains true for modern value-model architectures, regardless of whether they use supervised, unsupervised, or reinforcement learning.

A large literature treats pervasive tracking as presumptively objectionable [114, 4, 103]. If current recommenders require such tracking, we should explore designs that do not. It’s true that techniques like differential privacy, federated learning, and on-device inference can mask individual records. They do not remedy the collective and structural harms of mass surveillance—such as concentration of informational power or the chilling effects of ubiquitous monitoring [95, 4, 97, 87].

#### 3.2 Incentivizing Power Concentration

Existing recommender systems infer content representations, user preferences, and item-to-user match quality from behavioral data. The more such data they observe, the more accurate their models become. Crucially, obtaining that data at scale is feasible only when content distribution is centralized: platform owners must intermediate most user interactions so that every click, dwell, and share flows back into their training logs.

This feedback loop creates a clear strategic objective: maximize platform scale. More users and more content yield more data; more data improves recommendation quality; higher quality attracts still more users. Scale translates directly into concentrated power: the larger the platform, the more data is being collected, the more users’ attention is being allocated, the greater the degree of influence over people’s welfare, options, and attitudes [58] that sits with those who own the platform and its recommenders. The result: a small set of firms control the attention of billions, determining which signals are amplified and which are suppressed [32].

Philosophical critiques of such concentration of power over attention are well rehearsed [53, 94, 55], but recent events make the practical risks equally vivid. Entrusting the global flow of information to a handful of proprietors—sometimes idiosyncratic billionaires—invites arbitrary interventions in public discourse.

One response, middleware, would let third-party recommenders sit between the platform and the user [29]. In principle, a market of competing ranking services could break the present oligopoly over attention. In practice, middleware adoption has largely

stalled (notwithstanding the existence of ‘custom feeds’ on Bluesky) because behavioral data remain siloed. Platforms cite both competitive advantage and data-protection law when refusing to share those logs [52].

Hence, addressing power concentration requires recommender paradigms that do not depend on privileged access to platform-scale behavioral data. We turn to that possibility next.

### 3.3 Narrow Behaviorism

Current recommender systems infer user interests primarily from observable behavior—mouse-overs, scroll pauses, clicks, likes, reposts, and related signals [66, 100]. Although some contextual cues (e.g., platform assessments of content quality) are sometimes incorporated [18], behavioral data remain dominant. This emphasis produces narrowly behaviorist inferences that create at least two problems.

First, observed behavior tracks revealed rather than considered preferences [67]. Revealed preferences are the choices individuals make within the options they notice; considered preferences are the choices they would make with adequate information, a richer option set, and sufficient self-control. Because these conditions are often unmet [83], behavior is an unreliable proxy for considered preferences. For example, many users lament the heightened moral temperature of online discourse [84] yet still engage with outrage-inducing content [7, 8]. In the moment they may underestimate its cumulative costs, lack viable alternatives, or simply succumb to immediate impulses.

Second, a strictly behaviorist approach leaves scant room for encoding societal values [5]. These societal values extend beyond any one user’s short-term incentives and rarely manifest as discrete choices that a recommender can observe. We may, for instance, endorse a public sphere that supports democratic deliberation, yet be offered no individual action whose click-through meaningfully signals that commitment. Optimizing for short-run individual engagement therefore tends to amplify self-interested behavior and reproduce collective-action failures that most users would reject *ex ante*.

### 3.4 Compromised Attentional Agency

Current recommender systems are optimized to satisfy users’ immediate attentional impulses with minimal explicit input. They observe behavioral signals—pauses, mouse-overs, clicks, likes, reposts—and, by leveraging fast, automatic (‘System I’) cognitive processes [49], deliver a stream of content that can be consumed with little deliberation. Because these models learn almost exclusively from such passively collected traces, users cannot easily fine-tune them; meaningful ‘training’ requires extended usage [27] or coarse interventions such as muting keywords. Even where interface controls exist, adoption is typically limited to a small minority of users [48, 18]. The underlying high-dimensional representations are difficult to interpret, and external correction is currently infeasible [56].

These properties constrain *attentional agency*: the deliberate, reasons-responsive allocation of attention [105, 82]. While users *can* switch platforms or employ alternative discovery mechanisms, heavy reliance on opaque, hard-to-steer recommenders makes such agency more costly and so less frequent. Our claim is not that such systems eliminate agency altogether (see e.g. Hari [40]—but that they predictably weaken it).

Perceptions of weakened agency matter as well. When observers believe their interlocutors’ attention is mechanically directed by algorithms, they may attribute disagreement to manipulation rather than sincere difference of opinion, thereby fueling distrust [4].

In sum, contemporary recommenders incentivize large-scale behavioral surveillance, concentrate informational power, encode narrow behaviorism, and erode attentional agency. They are not the sole cause of dysfunction in the digital public sphere, but their

design choices incentivize specific, avoidable harms. Developing alternative approaches that achieve comparable utility without these costs is therefore a worthwhile research goal.

## 4 What Paths Forward?

Existing work suggests at least three responses to the shortcomings of algorithmic recommendation [110]. One is to abandon ML-based recommendation entirely. Another keeps the basic architecture of existing approaches, but aims to identify better behavioral signals. And we will introduce the third at the end of this section.

### 4.1 ‘Algorithms Ruin Everything’ [24]

A prominent reform proposal—championed on ‘federalist’ platforms such as Mastodon and Bluesky—eschews algorithmic recommendation in favor of a purely subscription-and social-network-driven model [10, 71]. Users encounter a post only if they have explicitly subscribed to its source or if it is reposted by an account they follow, with exposure mediated through their local social graph [26]. Items are then displayed in strict reverse-chronological order: an ordering rule that is simple and transparent, and does not predict relevance.

Although this design avoids certain pathologies of predictive ranking, dispensing with recommender algorithms altogether imposes substantial costs. First, it shifts the cognitive burden of discovery and prioritization onto each individual user—they demand *too much* attentional agency. Without automated ranking, feeds can become unmanageable as network size grows, and the only available controls are coarse, reactive measures (e.g., muting reposts or maintaining multiple curated lists). Second, the absence of relevance ordering can hinder real-time dialogue, especially across time-zone boundaries, because recent posts crowd out contextually salient—but slightly older—content.

In short, a feed determined solely by subscriptions preserves user autonomy but provides limited support for efficient discovery and attention management, especially the allocation of *collective* attention. A wholesale rejection of algorithmic recommendation therefore risks undermining the very communicative goals that social platforms are meant to serve.

### 4.2 Complementing Behaviorism

A more incremental strategy seeks to steer current recommender systems toward socially beneficial objectives by refining the metrics they optimize or the signals they incorporate [18]. Two avenues illustrate this approach:

- Richer explicit feedback. Expanding the range of user reactions—Facebook’s shift from a single ‘like’ to multiple emotional responses is a canonical example—supplies the algorithm with finer-grained preferences, giving users greater influence over their feeds.
- Proxy signals for civic value. Platforms can elevate content that performs well on metrics correlated with epistemic or social goods. X’s Community Notes, for instance, ranks annotations partly by whether they win approval from ideologically diverse reviewers, operationalizing a ‘bridging’ heuristic [75].

We do not attempt a comprehensive survey here, noting only two considerations. First, many such proposals merit serious experimentation. Second, as long as recommender systems remain grounded in large-scale behavioral surveillance and deep-learning inference controlled by centralized platforms, they will continue to incentivize data extraction,

concentrate power, exhibit elements of narrow behaviorism, and constrain user agency. Mitigating these deeper structural concerns may therefore require more radical departures from the current paradigm.

### 4.3 Language Models

Large language models have been used in recommendation tasks for a while as support for the heavy ranking tasks, mainly in the form of contextual embeddings for users' or items' text features. In recent times, however, their increased capabilities have opened new research directions, including the option to use such systems to perform ranking, as discussed in [22, 23, 104, 63, 111]. Bernstein et al. [5], Friedman et al. [28], Jia et al. [47], Lin et al. [62], Vats et al. [102], Huang et al. [44] have also considered how LLM-based recommenders can advance societal values specifically. Four capabilities of LLMs are especially salient to that task.

- **Functional language- and image-level understanding of content.** Irrespective of debates about ‘true’ semantic understanding [69], contemporary LLMs reliably classify content in accordance with rich evaluative criteria [112, 47, 54, 98, 45].
- **Better understanding of users** LLMs’ natural language competence also enables a new kind and degree of interaction with the user, allowing models to more deeply understand users’ preferences [61], as well as to provide explanations for any inferences made [30], something that traditional recommenders struggle with.
- **Reasoning and planning.** Although current models struggle in some distribution-shift scenarios [101, 60], they readily perform the core task of matching a content description to a preference profile and assigning a relevance score [5, 47, 48]. More complex pipelines embed models within agent frameworks that schedule multi-step plans [19, 31, 74, 21].
- **Tool use.** Via function-calling outputs, an LLM can invoke external software, evaluate the response, and incorporate the result into its context window for further processing [81, 43, 77]. More generally, this means that LLMs can operate as part of more complex systems, where the LLM is the executive center drawing on different software tools to enable it to achieve the tasks it has been set. Because these more complex systems involve LLMs undertaking complex sequences of action without supervision, based on its own reasoning capability, they are appropriately described as LLM agents [51].

These capacities support two broad system architectures:

**LM-assisted recommenders.** The LLM acts as a service that classifies items and scores user–item matches inside an otherwise conventional pipeline [5, 11].

**LM-agent recommenders.** The LLM holds executive control, deciding when to call auxiliary tools, retrieve new items, or update preference models (see figure above).

Regardless of architecture, an LLM-based recommender requires at least four modules. First, an inspectable, corrigible representation of each user’s preferences and relevant societal values. Second, a structured record of the user’s prior feed, capturing aggregate properties such as topical diversity and thematic recurrence. Third, a candidate-generation process—either platform-internal or agentic web-browsing. Fourth, a reasoning engine that combines (1)–(3) to rank items for delivery.

## 5 The Potential of an LLM-Enabled Public Sphere

The digital public sphere is dominated by firms whose business models reward large-scale behavioral surveillance and tight control over the flow of attention. If LLM-based recommenders are developed only inside those firms, they will probably inherit the same incentives [50]. Yet the capabilities of modern LLMs also make it technically possible to escape the four pathologies identified in Section 3. In this section we outline how.

## 5.1 Mitigating Mass Surveillance

LLM-based recommenders can evaluate candidate content and model user preferences directly through functional language and image understanding, and native reasoning capability, rather than through high-volume behavioral data and statistical inferences [5, 47]. Behavioral traces therefore become an avoidable externality rather than a fuel the system cannot run without. Users, or regulators, could simply forbid their collection without degrading ranking quality.

## 5.2 Diluting Concentrated Power

Because ranking quality depends mainly on the pretrained model rather than live network effects, LMAs do not automatically grow more accurate with every new user. Training frontier-scale models is still expensive, but three trends reduce the barrier to entry:

- Open-weights releases. Several near-frontier models are already available for commercial reuse.
- Distillation & synthetic data. Firms routinely shrink large checkpoints or bootstrap data from them, maintaining task performance in far smaller artifacts [14]
- Falling compute cost. Hardware and algorithmic gains [41, 17] are pushing frontier-level capability onto consumer devices; a two-year horizon for on-device operation of today’s most capable models is plausible.

Together these trends will make it feasible for individuals, research labs or small companies to run personal recommenders that are not beholden to platform gatekeepers.

In addition, LLM-based *agents* can act: they detect when external tools or browser calls are useful, execute them, evaluate the result and iterate [81, 31]. A personal agent can therefore discover material scattered across blogs, pre-print servers, mailing lists or proprietary platforms—even where APIs are throttled—then aggregate it into a single feed. For example, instead of the user having to check X because ‘that’s where the AI papers show up,’ an agent could crawl arXiv, RSS feeds, institutional repos and niche forums, returning a unified digest. Likewise, it could keep track of friends across Bluesky, Threads, LinkedIn and Mastodon, sparing users the pain of reverse-chronological doom-scrolling. Such bottom-up interoperability removes the audience lock-in on which contemporary network effects depend.<sup>2</sup>

On this model, platforms themselves would be likely to wither away (or else to radically change). And if everyone has their own agent advancing their own interests, then centralized control over the distribution of attention would also attenuate.

## 5.3 Beyond Narrow Behaviorism

LLMs can represent user and societal values in explicit natural-language form, accommodate higher-order preferences, and hold reflective dialog with the user [78]. Training an LLM-based agent recommender begins to resemble onboarding a human assistant: direct instructions lead to immediate generalization rather than months of implicit ‘training by scrolling.’ This also opens space for value-aligned nudging and for incorporating collective-risk constraints [5] instead of simply maximizing revealed preferences.

---

<sup>2</sup> In the EU the Digital Marketplace Act has an interoperability requirement, at present just for messaging apps. It has been agreed to discuss the feasibility of extending the requirement to social media too in future revisions. In the US the proposed ACCESS Act would also require interoperability between platforms.

## 5.4 Scaffolding Agency

Because the user model and inference chain can be surfaced in ordinary language, recommendations become explainable and corrigible. Users can ask: Why did you show me this?; edit the underlying preference statement; or override a specific suggestion. When well designed, the agent’s chain-of-thought (or equivalent structured rationale) is legible even if, at present, not perfectly faithful [99]. The net effect is augmented, not diminished, attentional agency.

LM-driven recommenders can mitigate surveillance, dilute platform power, defeat lock-in, transcend narrow behaviorism and enhance user agency. The next section analyses the practical, economic and normative hurdles that must be cleared to realise this potential.

## 6 Alternative Views: Challenges for LLM Recommenders

Large-language-model recommenders promise to reduce the harms associated with today’s engagement-optimized systems, but they introduce a distinct set of practical and ethical questions. We cannot resolve these issues *a priori*; the only decisive test is to prototype and evaluate full systems. Below we catalog the principal objections and outline why none is obviously fatal.

**Efficiency.** Even highly optimized LLMs on specialized hardware remain less efficient than recommender stacks refined over two decades.

Both algorithmic and hardware trends [41] point to rapid gains. Inference costs are already falling and are likely to approach those of traditional models within the product life-cycle of a new platform.

**Candidate Generation Requires Centralization.** Efficient first-stage retrieval appears to favor a single clearing-house that indexes content for everyone, recreating centralized power.

Retrieval can be hierarchical: lightweight local or edge models produce a coarse shortlist; stronger rerankers refine it. Falling compute costs and open-protocol infrastructure (e.g. the fediverse) mitigate the need for a single dominant index. Even if a single index were needed for candidate generation, this index need not replicate all the features of existing platforms, since agentic recommenders could absorb the other functions platforms currently perform. In general, novel infrastructure to support agents in general is needed, and can be designed around privacy and decentralization from the outset [50, 13].

**User Burden.** Systems that scaffold agency may impose unacceptable cognitive overhead; users prefer frictionless feeds.

There is no easy solution to this problem; we simply need to design UIs for LLM-based recommenders that judiciously balance user input with frictionless consumption.

**Value Representation.** LLMs have not yet shown stable, longitudinal modeling of individual preferences and values [89].

Ramos et al. [79] shows that systems that build natural-language preference profiles achieve performance comparable with matrix-factorization approaches; Zhou et al. [113] shows they can learn concise natural language preference profiles from conversation. Further progress may come from combining scaffolding and external tools to support structured generation and hierarchical representations of user values. For example, some preferences might be near-absolute constraints, some much softer. LLMs have made significant progress along this path, and promise more through the capabilities unlocked by scaling inference time compute [35, 73]. The more accurate these user models become, the more important it would be to ensure they are tightly controlled by users themselves, given the consequent privacy risk.

**Incumbent Resistance.** Platforms will adopt partial LLM solutions that preserve surveillance and power concentration, lobbying against open agents [50].

Given social media’s impact on the information and communication ecosystem, regulators clearly have motive and authority to ensure genuine competition in the allocation of attention, and to support LLM recommenders [15]. And more generally, with so many companies aiming to transform personal computing entirely through LLM agents [85], the rise of LLM recommenders might ultimately be a foregone conclusion. If LLM agents become the new operating system, and the universal interface for our digital lives, recommendation will likely just fold into that broader system. We will shift from a paradigm where digital platforms are the key intermediaries controlling our digital lives, to one where LLM agents play that role [57].

These *universal intermediaries* could allow power to shift from platforms to agents—but they could also be another vector for more fine-grained control of users’ lives by tech companies [50, 59].

**Security and Manipulation:** The security and safety of LLM agents raise ethical concerns. No current LLM system is immune to prompt injection [109]; hijacked agents could act on a user’s behalf, armed with a detailed personal model. Trusted agents may become potent vectors for manipulation, and decentralization complicates policing [12, 13].

The threat is serious but not clearly worse than today’s status quo. Robust security research, formal methods, capability bounding, and policy interventions are already under way and must be prioritized. Scaling inference compute also seems to promise enhanced adversarial robustness [108].

**Filter Bubbles:** LLM-based recommenders might create an acute version of the classic ‘filter bubble’ [92, 76].

This concern is likely to remain exaggerated. People can already choose to stay within their own narrow epistemic bubble, but most do not [1]. Agents can be steered toward diversity goals using the same natural-language constraints used for preference elicitation [5, 64]. Pathological designs can be forbidden by policy or protocol.

**Reduced Sociability:** Proxy agents might erode the public, interactive dimension of social media, turning it into a ‘ghost town’ of autonomous scrapers. Social media makes users’ tastes public, and supports social taste formation. We do not consume in isolation. Proxy agents would make this more challenging.

This is a design question, not an inevitability. We can specify when to delegate exploration to agents and when to surface direct human engagement. Nothing suggests these questions are intractable.

**Upstream Ethical Costs:** Some might argue against using LLMs for *anything*. Building with LLMs entails violating others’ copyright, ruining the environment, inescapably advancing the power of big tech, and so on (see e.g. [2] and the ensuing literature).

These critiques target training practices, not downstream use. Responsible data governance, green training pipelines, and diversified model supply can address the upstream harms. We reject the view that LLMs are so morally compromised as to preclude beneficial applications.

## 7 Conclusion

Early hopes that an open, networked web would democratize communication ultimately gave way to the concentrated power of a handful of platforms [3]. The emerging paradigm of LM-mediated personal computing could repeat that history; technological affordances alone cannot dissolve the structural inequities that shape today’s public sphere.

Yet new capabilities expand the design space. An architecture that enables richer local inference without pervasive surveillance, that moves beyond narrow behaviorist signals, and that treats users as active principals rather than passive optimization targets represents substantive progress over the status quo. The research agenda sketched here does not eliminate every risk, but it offers a credible path toward a healthier attention ecosystem—one worth pursuing despite the challenges ahead.

## References

- [1] Christopher Bail. *Breaking the social media prism : how to make our platforms less polarizing*. Princeton University Press, Princeton, 2021. ISBN 9780691203423.
- [2] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?, 2021. URL <https://doi.org/10.1145/3442188.3445922>.
- [3] Yochai Benkler. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press, 2006.
- [4] Claire Benn and Seth Lazar. What's wrong with Automated Influence. *Canadian Journal of Philosophy*, pages 1–24, 2021. doi: 10.1017/can.2021.23.
- [5] Michael Bernstein, Angèle Christin, Jeffrey Hancock, Tatsunori Hashimoto, Chenyan Jia, Michelle Lam, Nicole Meister, Nathaniel Persily, Tiziano Piccardi, and Martin Saveski. Embedding Societal Values into Social Media Algorithms. *Journal of Online Trust and Safety*, 2(1), 2023. ISSN 2770-3142.
- [6] Craig Boutilier, Martin Mladenov, and Guy Tennenholtz. Recommender Ecosystems: A Mechanism Design Perspective on Holistic Modeling and Optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20): 22575–22583, 2024. ISSN 2374-3468. doi: 10.1609/aaai.v38i20.30266. URL <https://ojs.aaai.org/index.php/AAAI/article/view/30266>.
- [7] William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318, 2017. ISSN 0027-8424.
- [8] William J. Brady, Killian L. McLoughlin, Mark P. Torres, Kara F. Luo, Maria Gendron, and M. J. Crockett. Overperception of moral outrage in online social networks inflates beliefs about intergroup hostility. *Nature Human Behaviour*, 7(6):917–927, 2023. ISSN 2397-3374. doi: 10.1038/s41562-023-01582-0. URL <https://doi.org/10.1038/s41562-023-01582-0>.
- [9] Emanuele Bugliarello, Rishabh Mehrotra, James Kirk, and Mounia Lalmas. Mostra: A Flexible Balancing Framework to Trade-off User, Artist and Platform Objectives for Music Sequencing. In *Proceedings of the ACM Web Conference 2022*, WWW '22, pages 2936–2945, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 978-1-4503-9096-5. doi: 10.1145/3485447.3512014. URL <https://doi.org/10.1145/3485447.3512014>.
- [10] Jenna Burrell, Zoe Kahn, Anne Jonas, and Daniel Griffin. When users control the algorithms: values expressed in practices on twitter. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–20, 2019. ISSN 2573-0142.
- [11] Diego Carraro and Derek Bridge. Enhancing Recommendation Diversity by Re-ranking with Large Language Models. *arXiv preprint*, page <https://arxiv.org/abs/2401.11506>, 2024.
- [12] Alan Chan, Noam Kolt, Peter Wills, Usman Anwar, Christian Schroeder de Witt, Nitarshan Rajkumar, Lewis Hammond, David Krueger, Lennart Heim, and Markus Anderljung. Ids for ai systems, 2024. URL <https://arxiv.org/abs/2406.12137>.

- [13] Alan Chan, Kevin Wei, Sihao Huang, Nitarshan Rajkumar, Elija Perrier, Seth Lazar, Gillian K. Hadfield, and Markus Anderljung. Infrastructure for ai agents, 2025. URL <https://arxiv.org/abs/2501.10114>.
- [14] Hailin Chen, Fangkai Jiao, Xingxuan Li, Chengwei Qin, Mathieu Ravaut, Ruochen Zhao, Caiming Xiong, and Shafiq Joty. ChatGPT’s One-year Anniversary: Are Open-Source Large Language Models Catching up? *arXiv preprint*, page <https://arxiv.org/abs/arXiv.2311.16989>, 2024. URL <http://arxiv.org/abs/2311.16989>.
- [15] Joshua Cohen and Archon Fung. *Democracy and the Digital Public Sphere*, pages 23–61. The University of Chicago Press, Chicago, 2021.
- [16] Rachel Cohen, Toby Newton-John, and Amy Slater. The relationship between Facebook and Instagram appearance-focused activities and body image concerns in young women. *Body image*, 23:183–187, 2017. ISSN 1740-1445.
- [17] Ben Cottier. Trends in the Dollar Training Cost of Machine Learning Systems. *Epoch AI*, 2023. URL <https://epochai.org/blog/trends-in-the-dollar-training-cost-of-machine-learning-systems>.
- [18] Tom Cunningham, Sana Pandey, Leif Sigerson, Jonathan Stray, Jeff Allen, Bonnie Barrilleaux, Ravi Iyer, Smitha Milli, Mohit Kothari, and Behnam Rezaei. What We Know About Using Non-Engagement Signals in Content Ranking. *arXiv preprint*, page <https://arXiv.org/abs/2402.06831>, 2024.
- [19] Gautier Dagan, Frank Keller, and Alex Lascarides. Dynamic planning with a llm. *arXiv preprint*, page <https://arXiv.org/abs/2308.06391>, 2023.
- [20] Fernando de la Cruz Paragas and Trisha TC Lin. Organizing and reframing technological determinism. *New Media & Society*, 18(8):1528–1546, 2016. doi: 10.1177/1461444814562156. URL <https://journals.sagepub.com/doi/abs/10.1177/1461444814562156>.
- [21] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghaio Lu, Shangyan Zhou, Shanhua Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen,

- Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL <https://arxiv.org/abs/2501.12948>.
- [22] Yashar Deldjoo, Zhankui He, Julian McAuley, Anton Korikov, Scott Sanner, Arnau Ramisa, René Vidal, Maheswaran Sathiamoorthy, Atoosa Kasirzadeh, and Silvia Milano. A review of modern recommender systems using generative models (gen-recsys). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6448–6458, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671474. URL <https://doi.org/10.1145/3637528.3671474>.
  - [23] Yashar Deldjoo, Zhankui He, Julian McAuley, Anton Korikov, Scott Sanner, Arnau Ramisa, Rene Vidal, Maheswaran Sathiamoorthy, Atoosa Kasirzadeh, Silvia Milano, and Francesco Ricci. Recommendation with generative models. *Foundations and Trends® in Information Retrieval*, pages 1–120, 09 2024.
  - [24] Michael A DeVito, Darren Gergle, and Jeremy Birnholtz. "Algorithms ruin everything" # RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 3163–3174, 2017.
  - [25] Elizabeth Dwoskin and Eugene Scott. Obama says tech companies have made democracy more vulnerable, 2022. URL <https://www.washingtonpost.com/technology/2022/04/21/obama-disinformation-stanford/>.
  - [26] David Easley and Jon Kleinberg. *Networks, crowds, and markets: reasoning about a highly connected world*. Cambridge University Press, New York, 2010. ISBN 9780511776755.
  - [27] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. First I" like" it, then I hide it: Folk Theories of Social Feeds. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2371–2382, 2016.
  - [28] Luke Friedman, Sameer Ahuja, David Allen, Terry Tan, Hakim Sidahmed, Changbo Long, Jun Xie, Gabriel Schubiner, Ajay Patel, and Harsh Lara. Leveraging Large Language Models in Conversational Recommender Systems. *arXiv preprint*, page <https://arxiv.org/abs/2305.07961>, 2023.
  - [29] Francis Fukuyama. Making the Internet Safe for Democracy. *Journal of Democracy*, 32(2):37–44, 2021.
  - [30] Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. Chat-rec: Towards interactive and explainable llms-augmented recommender system, 2023. URL <https://arxiv.org/abs/2303.14524>.

- [31] Zhi Gao, Yuntao Du, Xintong Zhang, Xiaojian Ma, Wenjuan Han, Song-Chun Zhu, and Qing Li. Clova: A closed-loop visual assistant with tool usage and update. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13258–13268, 2024.
- [32] Tarleton Gillespie. Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society*, 8(3):1–13, 2022. doi: 10.1177/20563051221117552. URL <https://journals.sagepub.com/doi/abs/10.1177/20563051221117552>.
- [33] Sandra González-Bailón, David Lazer, Pablo Barberá, Meiqing Zhang, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Deen Freelon, Matthew Gentzkow, Andrew M. Guess, Shanto Iyengar, Young Mie Kim, Neil Malhotra, Devra Moehler, Brendan Nyhan, Jennifer Pan, Carlos Velasco Rivera, Jaime Settle, Emily Thorson, Rebekah Tromble, Arjun Wilkins, Magdalena Wojcieszak, Chad Kiewiet de Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud, and Joshua A. Tucker. Asymmetric ideological segregation in exposure to political news on Facebook. *Science*, 381(6656):392–398, 2023. doi: doi:10.1126/science.adc7138. URL <https://www.science.org/doi/abs/10.1126/science.adc7138>.
- [34] Thore Graepel, Joaquin Quiñonero Candela, Thomas Borchert, and Ralf Herbrich. *Web-Scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft’s Bing search engine*, pages 13–20. Omnipress, Haifa, Israel, 2010.
- [35] Melody Y. Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, Hyung Won Chung, Sam Toyer, Johannes Heidecke, Alex Beutel, and Amelia Glaese. Deliberative alignment: Reasoning enables safer language models, 2025. URL <https://arxiv.org/abs/2412.16339>.
- [36] Andrew M. Guess and Benjamin A. Lyons. *Misinformation, Disinformation, and Online Propaganda*, pages 10–33. Cambridge University Press, Cambridge, 2020.
- [37] Andrew M. Guess, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, Sandra González-Bailón, Edward Kennedy, Young Mie Kim, David Lazer, Devra Moehler, Brendan Nyhan, Carlos Velasco Rivera, Jaime Settle, Daniel Robert Thomas, Emily Thorson, Rebekah Tromble, Arjun Wilkins, Magdalena Wojcieszak, Beixian Xiong, Chad Kiewiet de Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud, and Joshua A. Tucker. How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, 381(6656):398–404, 2023. doi: doi:10.1126/science.abp9364. URL <https://www.science.org/doi/abs/10.1126/science.abp9364>.
- [38] Jonathan Haidt and Christopher Bail. Social Media and Political Dysfunction: A Collaborative Review. *Unpublished MS.*, 2023.
- [39] Karen Hao. How Facebook got addicted to spreading misinformation, 2021. URL <https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/>.
- [40] Johann Hari. Stolen focus, 2021.
- [41] Anson Ho, Tamay Besiroglu, Ege Erdil, David Owen, Robi Rahman, Zifan Carl Guo, David Atkinson, Neil Thompson, and Jaime Sevilla. Algorithmic progress in language models. *arXiv preprint*, page <https://arxiv.org/abs/2403.05812>, 2024.

- [42] Erik Hoel. Here lies the internet, murdered by generative AI, 2024. URL <https://www.theintrinsicperspective.com/p/here-lies-the-internet-murdered-by>.
- [43] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=VtmBAGCN7o>.
- [44] Chengkai Huang, Junda Wu, Yu Xia, Zixu Yu, Ruhan Wang, Tong Yu, Ruiyi Zhang, Ryan A. Rossi, Branislav Kveton, Dongruo Zhou, Julian McAuley, and Lina Yao. Towards agentic recommender systems in the era of multimodal large language models, 2025. URL <https://arxiv.org/abs/2503.16734>.
- [45] Wenzuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models, 2025. URL <https://arxiv.org/abs/2503.06749>.
- [46] Dan Eilen Jasser Jasser and Ivan Garibay. Flooding the Zone: a censorship and disinformation strategy that needs attention. In *Proceedings of the 2022 International Workshop on Social Sensing (SocialSens 2022): Special Edition on Belief Dynamics*, 2021.
- [47] Chenyan Jia, Michelle S Lam, Minh Chau Mai, Jeff Hancock, and Michael S Bernstein. Embedding democratic values into social media AIs via societal objective functions. *arXiv preprint*, page <https://arxiv.org/abs/2307.13912>, 2023.
- [48] Tian Jin and Xin Dong. Recalign, 2023. URL <https://github.com/recalign/RecAlign>.
- [49] Daniel Kahneman. *Thinking, fast and slow*. Farrar, Straus and Giroux, New York, 2013. ISBN 9780374533557.
- [50] Sayash Kapoor, Noam Kolt, and Seth Lazar. Resist platform-controlled ai agents and champion user-centric agent advocates, 2025. URL <https://arxiv.org/abs/2505.04345>.
- [51] Atoosa Kasirzadeh and Iason Gabriel. Characterizing ai agents for alignment and governance, 2025. URL <https://arxiv.org/abs/2504.21848>.
- [52] Daphne Keller. The Future of Platform Power: Making Middleware Work. *Journal of Democracy*, 32(3):168–172, 2021. ISSN 1086-3214.
- [53] Daphne Keller. Amplification and its Discontents. *Knight First Amendment Institute at Columbia University Occasional Papers*, pages 1–47, 2021. URL <https://knightcolumbia.org/content/amplification-and-its-discontents>.
- [54] Joe Kwon, Josh Tenenbaum, and Sydney Levine. Neuro-Symbolic Models of Human Moral Judgment: LLMs as Automatic Feature Extractors. In *Proceedings of the 2023 ICML Workshop on Counterfactuals in Minds and Machines*, 2023.
- [55] Seth Lazar. Communicative Justice and the Distribution of Attention. *Knight First Amendment Institute*, pages

<http://knightcolumbia.tierraudev.com/content/communicative-justice-and-the-distribution-of-attention>, 2023. URL [https://knightcolumbia.org/content/communicative-justice-and-the-distribution-of-attention?\\_preview\\_=536fe16e10](https://knightcolumbia.org/content/communicative-justice-and-the-distribution-of-attention?_preview_=536fe16e10).

- [56] Seth Lazar. Legitimacy, Authority, and Democratic Duties of Explanation. *Oxford Studies in Political Philosophy*, 10:28–56, 2024.
- [57] Seth Lazar. Frontier AI Ethics: Anticipating and Evaluating the Societal Impacts of Language Model Agents. *arXiv preprint*, page <https://arxiv.org/abs/2404.06750>, 2024. URL <https://aeon.co/essays/can-philosophy-help-us-get-a-grip-on-the-consequences-of-ai>.
- [58] Seth Lazar. *Automatic Authorities: Power and AI*, page <https://arxiv.org/abs/2404.05990>. MIT Press, Cambridge, MA, 2024.
- [59] Seth Lazar. Governing the algorithmic city. *Philosophy & Public Affairs*, 53(2):102–168, 2025. doi: <https://doi.org/10.1111/papa.12279>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/papa.12279>.
- [60] Martha Lewis and Melanie Mitchell. Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models. *arXiv preprint*, page <https://arxiv.org/abs/2402.08955>, 2024.
- [61] Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. Eliciting human preferences with language models, 2023. URL <https://arxiv.org/abs/2310.11589>.
- [62] Jianghao Lin, Xinyi Dai, Yunjia Xi, Weiwen Liu, Bo Chen, Hao Zhang, Yong Liu, Chuhan Wu, Xiangyang Li, and Chenxu Zhu. How can recommender systems benefit from large language models: A survey. *arXiv preprint*, page <https://arxiv.org/abs/2306.05817>, 2023.
- [63] Jianghao Lin, Xinyi Dai, Yunjia Xi, Weiwen Liu, Bo Chen, Hao Zhang, Yong Liu, Chuhan Wu, Xiangyang Li, Chenxu Zhu, Huifeng Guo, Yong Yu, Ruiming Tang, and Weinan Zhang. How can recommender systems benefit from large language models: A survey, 2024. URL <https://arxiv.org/abs/2306.05817>.
- [64] Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M Dai, Diyi Yang, and Soroush Vosoughi. Training socially aligned language models on simulated social interactions. *The Twelfth International Conference on Learning Representations*, page <https://openreview.net/forum?id=NddKiWtdUm>, 2024.
- [65] Meta. Introducing 22 system cards that explain how AI powers experiences on Facebook and Instagram. *Meta AI*, pages <https://ai.meta.com/blog/how-ai-powers-experiences-facebook-instagram-system-cards/>, 6 2023. URL <https://ai.meta.com/blog/how-ai-powers-experiences-facebook-instagram-system-cards/>.
- [66] Meyerson, Eric. YouTube Now: Why We Focus on Watch Time. *blog.youtube*, pages <https://blog.youtube/news-and-events/youtube-now-why-we-focus-on-watch-time/>, 8 2012. URL <https://blog.youtube/news-and-events/youtube-now-why-we-focus-on-watch-time/>.
- [67] Smitha Milli, Luca Belli, and Moritz Hardt. From Optimizing Engagement to Measuring Value, 2021.

- [68] Smitha Milli, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao, and Anca D Dragan. Engagement, user satisfaction, and the amplification of divisive content on social media. *arXiv preprint*, page <https://arxiv.org/abs/2305.16941>, 2023.
- [69] Jared Moore. Language Models Understand Us, Poorly. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 214–222. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-emnlp.16. URL <https://doi.org/10.18653/v1/2022.findings-emnlp.16>.
- [70] Kevin Munger and Joseph Phillips. Right-Wing YouTube: A Supply and Demand Perspective. *The International Journal of Press/Politics*, 27(1):186–219, 2020. doi: 10.1177/1940161220964767. URL <https://journals.sagepub.com/doi/abs/10.1177/1940161220964767>.
- [71] Arvind Narayanan. Understanding Social Media Recommendation Algorithms. *Knight First Amendment Institute*, pages 1–49, 2023.
- [72] Brendan Nyhan, Jaime Settle, Emily Thorson, Magdalena Wojcieszak, Pablo Barberá, Annie Y. Chen, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, Sandra González-Bailón, Andrew M. Guess, Edward Kennedy, Young Mie Kim, David Lazer, Neil Malhotra, Devra Moehler, Jennifer Pan, Daniel Robert Thomas, Rebekah Tromble, Carlos Velasco Rivera, Arjun Wilkins, Beixian Xiong, Chad Kiewiet de Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud, and Joshua A. Tucker. Like-minded sources on Facebook are prevalent but not polarizing. *Nature*, 620(7972):137–144, 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06297-w. URL <https://doi.org/10.1038/s41586-023-06297-w>.
- [73] OpenAI. OpenAI o3 and o4-mini System Card, 4 2025. URL <https://openai.com/index/o3-o4-mini-system-card/>. Accessed: 2025-05-21.
- [74] OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Rasó, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O’Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu,

Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madeleine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2024. URL <https://arxiv.org/abs/2412.16720>.

- [75] Aviv Ovadya and Luke Thorburn. Bridging Systems: Open Problems for Countering Destructive Divisiveness Across Ranking, Recommenders and Governance. *Knight First Amendment Institute*, 2023. URL <https://knightcolumbia.org/content/bridging-systems>.
- [76] Eli Pariser. *The filter bubble : what the Internet is hiding from you*. Viking, London, 2011. ISBN 067092038X.
- [77] Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, Wanjun Zhong, Kuanye Li, Jiale Yang, Yu Miao, Woyu Lin, Longxiang Liu, Xu Jiang, Qianli Ma, Jingyu Li, Xiaojun Xiao, Kai Cai, Chuang Li, Yaowei Zheng, Chaolin Jin, Chen Li, Xiao Zhou, Minchao Wang, Haoli Chen, Zhaojian Li, Haihua Yang, Haifeng Liu, Feng Lin, Tao Peng, Xin Liu, and Guang Shi. Ui-tars: Pioneering automated gui interaction with native agents, 2025. URL <https://arxiv.org/abs/2501.12326>.
- [78] Filip Radlinski, Krisztian Balog, Fernando Diaz, Lucas Dixon, and Ben Wedin. On natural language user profiles for transparent and scrutable recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2863–2874, 2022.
- [79] Jerome Ramos, Hossen A Rahmani, Xi Wang, Xiao Fu, and Aldo Lipani. Natural Language User Profiles for Transparent and Scrutable Recommendations. *arXiv preprint*, page <https://arxiv.org/abs/2402.05810>, 2024.

- [80] Nardine Saad. Jon Stewart calls ‘overreaction’ to Spotify’s Joe Rogan debacle ‘a mistake’, 2022. URL <https://www.latimes.com/entertainment-arts/tv/story/2022-02-04/jon-stewart-joe-rogan-spotify-neil-young-overreaction>.
- [81] Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Tool-former: Language Models Can Teach Themselves to Use Tools. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 68539–68551. Curran Associates, Inc., 2023. URL [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/d842425e4bf79ba039352da0f658a906-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/d842425e4bf79ba039352da0f658a906-Paper-Conference.pdf).
- [82] Nick Schuster and Seth Lazar. Attention, moral skill, and algorithmic recommendation. *Philosophical Studies*, pages 1–26, 2024. ISSN 0031-8116.
- [83] Amartya Sen. Utilitarianism and Welfarism. *The Journal of Philosophy*, 76(9):463–489, 1979. URL <http://links.jstor.org/sici?&sici=0022-362X%28197909%2976%3A9%3C463%3AUAW%3E2.0.CO%3B2-D>.
- [84] Jaime E. Settle. *Frenemies: How Social Media Polarizes America*. Cambridge University Press, 2018.
- [85] Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O’Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, and Alan Hickey. Practices for governing agentic AI systems. Report, OpenAI, 2023.
- [86] Alexandra A. Siegel. *Online Hate Speech*, pages 56–88. Cambridge University Press, Cambridge, 2020. ISBN 9781108835558. doi: DOI:10.1017/9781108890960.
- [87] Mary Anne Smart, Dhruv Sood, and Kristen Vaccaro. Understanding Risks of Privacy Theater with Differential Privacy. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2):Article 342, 2022. doi: 10.1145/3555762. URL <https://doi.org/10.1145/3555762>.
- [88] Ben Smith. How TikTok Reads Your Mind, 2021. URL <https://www.nytimes.com/2021/12/05/business/media/tiktok-algorithm.html>.
- [89] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofer Mireshghalah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, and Nouha Dziri. A Roadmap to Pluralistic Alignment. *arXiv preprint*, page <https://arxiv.org/abs/2402.05070>, 2024.
- [90] Tom Standage. *Writing on the wall: social media - the first 2,000 years*. Bloomsbury, London, 2013. ISBN 9781408842065 (pbk.).
- [91] Galen Stocking, Amy Mitchell, Katerina Eva Matsa, Regina Widjaya, Mark Jurkowitz, Shreenita Ghosh, Aaron Smith, Sarah Naseer, and Christopher St. Aubin. The Role of Alternative Social Media in the News and Information Environment. Report, Pew Research Center, 2022.
- [92] Cass R. Sunstein. *Republic.com*. Princeton University Press, Princeton, N.J. ; Oxford, 2001. ISBN 0691070253 (alk. paper).
- [93] Daniel Susser, Beate Roessler, and Helen Nissenbaum. Online manipulation: Hidden influences in a digital world. *Georgetown Law Technology Review*, 4:1–45, 2019.

- [94] Jamie Susskind. *The Digital Republic: On freedom and democracy in the 21st century*. Bloomsbury Publishing, 2022.
- [95] Linnet Taylor, Luciano Floridi, and Bart van der Sloot. *Group Privacy: New Challenges of Data Technologies*. Philosophical Studies Series. Springer, 2017.
- [96] Luke Thorburn, Priyanjana Bengani, and Jonathan Stray. How Platform Recommenders Work, 2022. URL <https://medium.com/understanding-recommenders/how-platform-recommenders-work-15e260d9a15a>.
- [97] Martin Tisné. Collective data rights can stop big tech from obliterating privacy, 2021. URL <https://www.technologyreview.com/2021/05/25/1025297/collective-data-rights-big-tech-privacy/>.
- [98] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. *arXiv preprint arXiv:2406.16860*, 2024.
- [99] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language Models Dont Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 74952–74965. Curran Associates, Inc., 2023. URL [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/ed3fea9033a80fea1376299fa7863f4a-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/ed3fea9033a80fea1376299fa7863f4a-Paper-Conference.pdf).
- [100] Twitter. `twitter/the-algorithm`, 2024. URL <https://github.com/twitter/the-algorithm>.
- [101] Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. On the Planning Abilities of Large Language Models - A Critical Investigation. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 75993–76005. Curran Associates, Inc., 2023. URL [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/efb2072a358cef75886a315a6fcf880-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/efb2072a358cef75886a315a6fcf880-Paper-Conference.pdf).
- [102] Arpita Vats, Vinija Jain, Rahul Raja, and Aman Chadha. Exploring the Impact of Large Language Models on Recommender Systems: An Extensive Review. *arXiv preprint arXiv:2402.18590*, 2024.
- [103] Carissa Véliz. *Privacy is power : Why and how you should take back control of your data*. Penguin Books, London, 2021. ISBN 9780552177719 (pbk.).
- [104] Qi Wang, Jindong Li, Shiqi Wang, Qianli Xing, Runliang Niu, He Kong, Rui Li, Guodong Long, Yi Chang, and Chengqi Zhang. Towards next-generation llm-based recommender systems: A survey and beyond, 2024. URL <https://arxiv.org/abs/2410.19744>.
- [105] James Williams. *Stand out of our light: freedom and resistance in the attention economy*. Cambridge University Press, Cambridge, 2018. ISBN 9781108429092.
- [106] Samuel C. Woolley. *Bots and Computational Propaganda: Automation for Communication and Control*, pages 89–110. Cambridge University Press, Cambridge, 2020.

- [107] Tim Wu. Is the First Amendment Obsolete? *Michigan law review*, 117(3):547, 2018. ISSN 0026-2234. doi: 10.36644/mlr.117.3.first.
- [108] Wojciech Zaremba, Evgenia Nitishinskaya, Boaz Barak, Stephanie Lin, Sam Toyer, Yaodong Yu, Rachel Dias, Eric Wallace, Kai Xiao, Johannes Heidecke, and Amelia Glaese. Trading inference-time compute for adversarial robustness, 2025. URL <https://arxiv.org/abs/2501.18841>.
- [109] Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. InjecAgent: Benchmarking Indirect Prompt Injections in Tool-Integrated Large Language Model Agents. *arXiv preprint*, page <https://arxiv.org/abs/2403.02691>, 2024.
- [110] Amy X Zhang, Michael S Bernstein, David R Karger, and Mark S Ackerman. Form-From: A Design Space of Social Media Systems. *arXiv preprint*, page <https://arxiv.org/abs/2402.05388>, 2024.
- [111] Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, and Qing Li. Recommender systems in the era of large language models (llms). *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6889–6907, November 2024. ISSN 2326-3865. doi: 10.1109/tkde.2024.3392335. URL <http://dx.doi.org/10.1109/TKDE.2024.3392335>.
- [112] Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. GPT-4V(ision) is a Generalist Web Agent, if Grounded. *arXiv preprint*, page <https://arxiv.org/abs/2401.01614>, 2024.
- [113] Joyce Zhou, Yijia Dai, and Thorsten Joachims. Language-Based User Profiles for Recommendation. *arXiv preprint*, page <https://arxiv.org/abs/2402.15623>, 2024.
- [114] Shoshana Zuboff. *The Age of Surveillance Capitalism*. Public Affairs, New York, 2019.